On Collapsed State-space Models and the Particle Marginal Metropolis-Hastings Sampler

Lawrence M. Murray

CSIRO Mathematics, Informatics and Statistics
Perth WA, Australia

Emlyn Jones and John Parslow

CSIRO Marine and Atmospheric Research
Hobart Tasmania, Australia

February 29, 2012

Abstract

Monte Carlo sampling of nonlinear state-space models is particularly difficult in circumstances where the transition density is not of closed form. This is common in the physical and biological sciences, where real phenomena demand modelling that accurately reproduces nonlinearity, chaos, mass conservation, multiple potentials and other behaviours that do not necessarily yield convenient analytical forms. In this work, we explore a new idea in the design of samplers for such models, collapsing out the state variables of a state-space model to leave only its noise terms. We then consider the design of proposal distributions over these noise terms, exploiting the independence and simple parametric forms prescribed to them in the prior structure of the model.

The chosen vehicle for joint state and parameter estimation in these collapsed state-space models is the particle marginal Metropolis-Hastings (PMMH) sampler, from the family of particle Markov chain Monte Carlo (PMCMC). An auxiliary particle filter is used in the inner loop, with proposal distributions over noise terms tuned using the unscented Kalman filter. We conduct a thorough empirical investigation of the PMMH sampler in this context, including a look at the improvement obtained in acceptance rates, convergence rates and variability of the likelihood estimator. Case studies from the domain of marine biogeochemistry are used to demonstrate the ideas. We believe that these cases, exhibiting mass conservation and mild chaotic behaviour, are particularly challenging applications on which to exercise the PMCMC methodology.

Keywords: marine biogeochemistry, particle filter, particle Markov chain Monte Carlo, sequential Monte Carlo, unscented Kalman filter

1 Introduction

The tuning of a proposal distribution to improve the performance properties of a Monte Carlo sampler requires both a prior density and likelihood function that are computationally tractable. In rejection, importance and Markov chain Monte Carlo (MCMC) sampling, computations appear of the form:

$$\frac{\text{likelihood} \times \text{prior density}}{\text{proposal density}}\,. \tag{1}$$

In the case of a state-space model (detailed further in §2), the prior density is:

$$p(\mathbf{x}_{0:T}, \boldsymbol{\theta}) = p(\mathbf{x}_0|\boldsymbol{\theta})p(\boldsymbol{\theta}) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\theta})\,, \tag{2}$$

and is usually specified via explicit densities $p(\mathbf{x}_0|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$, with the remainder implied by the *transition density* $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\theta})$. If the transition density has a convenient closed form, such as for conditional Gaussian models, then the prior distribution is computationally tractable. This facilitates sampling using arbitrary proposal distributions. Across the spectrum of potential applications for nonlinear state-space models, however, such cases may well be exceptional, not the common norm. If the transition distribution can be simulated, but has no closed-form probability density function [see e.g. 10, 8, 24], then the prior distribution is computationally intractable. In such cases, one approach is to constrain the candidature of proposal distributions to those that cancel the prior density (or at least its intractable factors) in (1). Simulation from the prior must feature in the proposal to achieve this, as in the bootstrap particle filter [11]. An alternative is to unbiasedly estimate the transition density, as in the random-weight particle filter [8]. Both approaches can be quite limiting in a multivariate setting.

Models without closed-form transition densities are common in the biological and physical sciences, where nonlinear functional forms of increasing complexity are used to capture various phenomena in ever-greater detail. Such models may feature bifurcation, multi-well potentials or chaos, and stochasticity may enter by continuous, currency conserving or other elaborate means to render unwieldy transition densities.

The first contribution of this work is to remedy the proposal problem by collapsing the state-space model to its noise terms: the independent, univariate terms of prescribed distribution that form the source of stochasticity (Figure 1). This can be done for any model, or at least, pragmatically, must be doable if we intend to rely on sequences of independent (pseudo-)random numbers. The effect is that the likelihood function subsumes the nonlinear complexity of the model, leaving a simple prior density over noise terms. The transition density becomes trivial, noise terms being independent *a priori*, and one can begin to design proposal distributions over these. We should caution that the design of such proposals may be difficult, a strongly nonlinear mapping between noise terms and observations potentially producing complex posterior distributions over those noise terms, which may be difficult to approximate. Where a closed-form transition density is available, collapsing to noise terms is unlikely to be of benefit. Nevertheless, one might appreciate the ability to exert at least

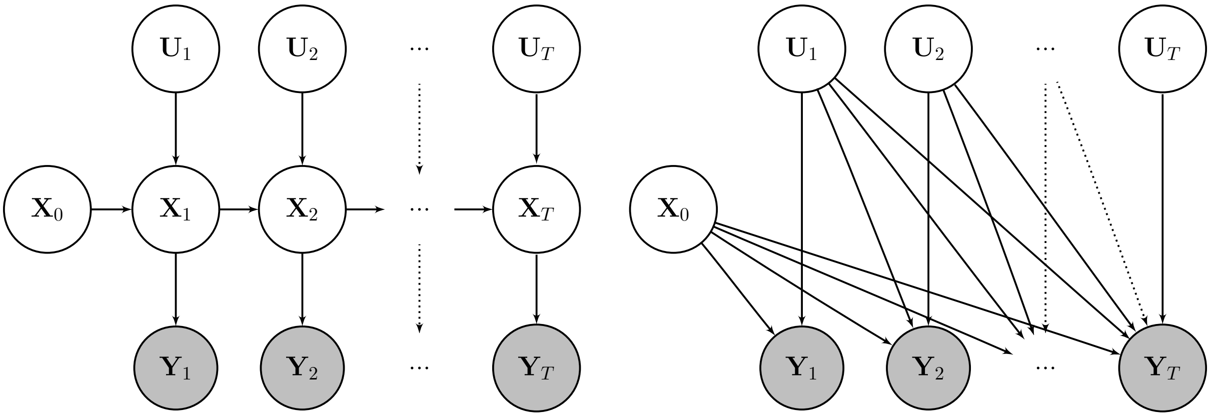

Figure 1: Graphical models of **(a)** conventional state-space model, and **(b)** collapsed state-space model, where the conventional model is reduced to initial conditions, noise terms and observations. Parameters, $\mathbf{\Theta}$, have been removed for clarity, but note that $\mathbf{U}_{1:T} \perp\!\!\!\perp \mathbf{\Theta}$, while all other variables depend on them.

some control, however limited, for a class of models where inference procedures are usually quite constrained.

By way of context and completeness, an intractable likelihood function, rather than prior density, would lend itself to Approximate Bayesian Computation [29]. The proposed collapse to noise terms differs from the separation of tractable and intractable components in the Rao-Blackwellisation of state-space models [5], where it is typical to admit dependence of the tractable (linear) component on the intractable (nonlinear) component, but not vice-versa; in the collapsed state-space model described here, the intractable (nonlinear) component depends on the tractable (independent noise) component.

In this work, the particle marginal Metropolis-Hastings (PMMH) sampler is employed, a particular particle Markov chain Monte Carlo (PMCMC) algorithm [1]. The unscented Kalman filter [UKF, 15, 39] is used to construct the proposal distributions via which improved performance of that PMMH sampler is sought. Two models from the domain of marine biogeochemistry are used as a study for this, previously treated with simpler PMMH approaches in Jones et al. [14] and Parslow et al. [25]. To tame the PMMH sampler, we must better understand its behaviour in the wild: quite some space is dedicated to this, and the insight and diagnostics developed form the second major contribution of the work.

The collapsed state-space model is described in §2. PMMH, along with preliminaries on its behaviour, are described in §3. Particular proposal mechanisms within PMMH using the UKF are given in §4. Case studies feature in §5, with detailed discussion and conclusions in §6. Appendices pick up the statistic used for assessing the likelihood estimator across space (§A), efficiencies for the special case of models with only Gaussian noise (§B), along with other non-essential implementation issues, and a description of the substantive model used in the second case study (§C).

2 The collapsed state-space model

For T time points, a sequence of observations $\mathbf{y}_1, \ldots, \mathbf{y}_T$ of random variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_T \in \mathbb{R}^{N_y}$ is assumed given, indicative of latent states $\mathbf{X}_1, \ldots, \mathbf{X}_T \in \mathbb{R}^{N_x}$ and initial condition $\mathbf{X}_0 \in \mathbb{R}^{N_x}$. The model is parameterised by static $\boldsymbol{\Theta} \in \mathbb{R}^{N_\theta}$, with the hierarchical model:

$$p(\mathbf{y}_{1:T}, \mathbf{x}_{0:T}, \boldsymbol{\theta}) = \left[\prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta})\right] \left[\prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\theta})\right] p(\mathbf{x}_0|\boldsymbol{\theta})p(\boldsymbol{\theta}) \,. \tag{3}$$

We follow the convention that uppercase letters denote random variables, while lowercase letters denote values of them. Bold typeface indicates vectors. For some random variable Λ, the notation $\Lambda_{s:t}$ denotes the sequence $\Lambda_s, \Lambda_{s+1}, \ldots, \Lambda_t$, and $\lambda_t^{m:n}$ the sequence of Λ_t samples, $\lambda_t^m, \lambda_t^{m+1}, \ldots, \lambda_t^n$.

This standard state-space model formulation is depicted in Figure 1(a), along with the independent noise terms $\mathbf{U}_1, \ldots, \mathbf{U}_T \in \mathbb{R}^{N_u}$ that drive the transition. In the context of the models under study, it is expedient to collapse the latent state to these noise terms only, as in Figure 1(b), where the $\mathbf{X}_{1:T}$ are eliminated as mere intermediates in the mapping from $\{\mathbf{X}_0, \mathbf{U}_{1:T}, \boldsymbol{\Theta}\}$ to $\{\mathbf{Y}_{1:T}\}$. This permits the design of samplers over the random variables $\{\mathbf{X}_0, \mathbf{U}_{1:T}, \boldsymbol{\Theta}\}$, rather than the typical set $\{\mathbf{X}_{0:T}, \boldsymbol{\Theta}\}$. A closed-form transition density is assured by the independence prescribed to $\mathbf{U}_{1:T}$ in the model's prior structure. In this work, the $\mathbf{U}_{1:T}$ are always standard i.i.d. $\mathcal{N}(0,1)$ variates, independent of parameters in the prior structure, but scaled and translated by them as required. Variates of any base distribution might be equally used, however. The new hierarchical model is

$$p(\mathbf{y}_{1:T}, \mathbf{u}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}) = \left[\prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{u}_{1:t}, \mathbf{x}_0, \boldsymbol{\theta})\right] \left[\prod_{t=1}^{T} p(\mathbf{u}_t)\right] p(\mathbf{x}_0|\boldsymbol{\theta})p(\boldsymbol{\theta}) \,. \tag{4}$$

Simulation of the process model is described by the function $f : \{\mathbf{U}_t, \mathbf{X}_{t-1}, \boldsymbol{\Theta}\} \to \mathbf{X}_t$. Recursive application of $f(\cdot)$ can be used to recover the state trajectory $\mathbf{x}_{1:t}$ from any sample $\{\mathbf{u}_{1:t}, \mathbf{x}_0, \boldsymbol{\theta}\}$.

For simplicity, we assume that N_x, N_u and N_y are constant across time. Note, however, that the methods do not depend on this, or even on equispaced observations; they are equally applicable with changing state and observation sizes, such as under missing or sparse data.

3 Inference methods for the collapsed state-space model

The target (posterior) density is $\pi(\mathbf{u}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}|\mathbf{y}_{1:T})$, factorised as either:

$$\pi(\mathbf{u}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}|\mathbf{y}_{1:T}) \quad = \quad \pi_1(\mathbf{u}_{1:T}, \mathbf{x}_0|\boldsymbol{\theta}, \mathbf{y}_{1:T})\pi_2(\boldsymbol{\theta}|\mathbf{y}_{1:T}) \,. \tag{5}$$

or

$$\pi(\mathbf{u}_{1:T}, \mathbf{x}_0, \boldsymbol{\theta}|\mathbf{y}_{1:T}) \quad = \quad \pi_1(\mathbf{u}_{1:T}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:T})\pi_2(\mathbf{x}_0, \boldsymbol{\theta}|\mathbf{y}_{1:T}) \,. \tag{6}$$

In either case, the first factor, $\pi_1(\cdot)$, is targeted using the formulation of the auxiliary particle filter [APF, 26], described in §3.1. The second factor, $\pi_2(\cdot)$, is targeted in an outer loop around the particle filter using Metropolis-Hastings [MH, 21, 12]. The particle filter nested within MH defines particle marginal Metropolis-Hastings [PMMH, 1], described in §3.2.

The factorisation (6) is attractive over (5) in circumstances where prior information over initial conditions is scarce, and an importance sample over them would be degenerate. It is used for the derivations here; use of the alternative is straightforward.

3.1 Auxiliary particle filtering

For given $\{\mathbf{x}_0, \boldsymbol{\theta}\}$, estimation of $\pi_1(\mathbf{u}_{1:t}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$ may proceed recursively as:

$$\begin{aligned} p(\mathbf{u}_{1:t}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t}) &\propto p(\mathbf{y}_t|\mathbf{u}_{1:t}, \mathbf{x}_0, \boldsymbol{\theta})p(\mathbf{u}_{1:t}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1}) & (7)\\ &= p(\mathbf{y}_t|\mathbf{u}_{1:t}, \mathbf{x}_0, \boldsymbol{\theta})p(\mathbf{u}_t|\mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})p(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1}) & (8)\\ &= p(\mathbf{y}_t|\mathbf{u}_{1:t}, \mathbf{x}_0, \boldsymbol{\theta})p(\mathbf{u}_t)p(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})\,. & (9) \end{aligned}$$

The key step, and motivation for operating over $\{\mathbf{U}_{1:T}, \mathbf{X}_0, \boldsymbol{\Theta}\}$, is the third. An orthodox derivation over $\{\mathbf{X}_{0:T}, \boldsymbol{\Theta}\}$ yields the transition density $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\theta})$, for which a closed form is not expected in the models under study. The collapsed state-space model instead produces the term $p(\mathbf{u}_t)$, of closed form by definition.

At time t, a sequential importance sampling of (9) proceeds with a draw $\mathbf{u}_{1:t}^m \sim q_t(\mathbf{u}_{1:t})$, weighted with

$$w_t^m = \frac{p(\mathbf{y}_t|\mathbf{u}_{1:t}^m, \mathbf{x}_0, \boldsymbol{\theta})p(\mathbf{u}_t^m)p(\mathbf{u}_{1:t-1}^m|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})}{q_t(\mathbf{u}_{1:t}^m)}\,. \tag{10}$$

Here, $q_t(\mathbf{u}_{1:t})$ is some proposal distribution satisfying the usual assumptions for sequential importance sampling [4].

The reduction to noise terms does not produce a method profoundly different to the conventional. The construction is simply that of the auxiliary particle filter [26], reduced to $\{\mathbf{U}_{1:T}, \mathbf{X}_0, \boldsymbol{\Theta}\}$. State samples $\mathbf{x}_t^m = f(\mathbf{u}_t^m, \mathbf{x}_{t-1}^m, \boldsymbol{\theta})$ may be stored and recursively updated in place of complete trajectories of noise terms $\mathbf{u}_{1:t}^m$, and no densities that would require a Jacobian correction for change-of-variable are ever explicitly evaluated. The only change is to enable arbitrary proposal distributions in the space of noise terms.

A fully adapted [27, 28] proposal would take the form

$$q_t(\mathbf{u}_{1:t}) = q_1(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})q_2(\mathbf{u}_t|\mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})\,. \tag{11}$$

It is not expected that full adaptation can be obtained in general, but reduction to noise terms does facilitate partial adaptation – various approximate substitutes for $q_1(\cdot)$ and $q_2(\cdot)$. Sampling from $q_1(\cdot)$ typically involves performing some auxiliary *look-ahead* procedure to time t [18], yielding weights $\omega_t^{1:M}$, from which a *look-behind* resample at time $t-1$ is performed to produce samples. The resampled particles are then propagated forward, with proposals drawn from $q_2(\cdot)$. We refer to the weights $\omega_t^{1:M}$ as *stage-one* weights, and the final weights $w_t^{1:M}$ as *stage-two* weights. The concepts

of look-ahead and look-behind are purposely distinguished here to avoid confusion later: some methods will feature a look-ahead using a UKF to construct $q_2(\cdot)$, without then using a look-behind to construct $q_1(\cdot)$.

Pitt [27] considers estimation of the likelihood $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$ with the APF, equally applicable to $p(\mathbf{y}_{1:T}|\mathbf{x}_0, \boldsymbol{\theta})$. The estimator uses both stage-one and stage-two weights, and is given by

$$p(\mathbf{y}_{1:T}|\mathbf{x}_0, \boldsymbol{\theta}) \approx \prod_{t=1}^{T} \left\{ \frac{1}{M} \sum_{m=1}^{M} w_t^m \right\} \left\{ \sum_{m=1}^{M} \omega_t^m \right\} . \quad (12)$$

Proofs of unbiasedness are given in Del Moral [4] and Pitt et al. [28].

3.2 The particle marginal Metropolis-Hastings sampler

At a high-level, the PMMH sampler combines a MH sampling over $\{\mathbf{X}_0, \boldsymbol{\Theta}\}$ with a sequential Monte Carlo (SMC) sampling over $\{\mathbf{X}_{1:T}|\mathbf{X}_0, \boldsymbol{\Theta}\}$, or $\{\mathbf{U}_{1:T}|\mathbf{X}_0, \boldsymbol{\Theta}\}$ in the collapsed case. In the outer loop, a proposed move $\{\mathbf{x}_0, \boldsymbol{\theta}\} \to \{\mathbf{x}'_0, \boldsymbol{\theta}'\} \sim q(\mathbf{x}'_0, \boldsymbol{\theta}'|\mathbf{x}_0, \boldsymbol{\theta})$ is accepted with probability

$$\min \left[1, \frac{p(\mathbf{y}_{1:T}|\mathbf{x}'_0, \boldsymbol{\theta}')p(\mathbf{x}'_0, \boldsymbol{\theta}')q(\mathbf{x}_0, \boldsymbol{\theta}|\mathbf{x}'_0, \boldsymbol{\theta}')}{p(\mathbf{y}_{1:T}|\mathbf{x}_0, \boldsymbol{\theta})p(\mathbf{x}_0, \boldsymbol{\theta})q(\mathbf{x}'_0, \boldsymbol{\theta}'|\mathbf{x}_0, \boldsymbol{\theta})} \right] , \quad (13)$$

where the likelihoods are estimated by SMC targeting $\pi_1(\cdot)$ in (6); in this work by an APF as described above.

Rigorously, $\pi_1(\cdot)$ is a marginal of a distribution over an extended space in which the PMCMC algorithm operates, a space which includes variables associated with the resampling mechanism of the particle filter [c.f. Equation 22 of 1]:

$$\psi(\mathbf{u}_{1:T}^{1:M}, a_{1:T}^{1:M} \mid \mathbf{x}_0, \boldsymbol{\theta}) = \prod_{t=1}^{T} \left[r(a_t^{1:M} \mid w_{t-1}^{1:M}) \prod_{p=1}^{M} p(\mathbf{u}_t^m \mid \mathbf{u}_1^{b_1^m}, \ldots, \mathbf{u}_{t-1}^{b_{t-1}^m}, \mathbf{x}_0, \boldsymbol{\theta}) \right] . \quad (14)$$

Here, a_t^m is the index of the particle at time $t-1$ which is the ancestor of particle m at time t, as selected by the resampling algorithm. By recursively tracing these indices backward to time s, where $s \leq t$, the notation b_s^m denotes the index of the particle at time s which is the ancestor of particle m at time t. For some specific $m \sim \mathrm{Multinomial}(w_T^{1:M})$, a sample of $\pi_1(\cdot)$ is given by the marginal $\{\mathbf{u}_1^{b_1^m}, \ldots, \mathbf{u}_T^{b_T^m}\}$, the precise form of which is not required.

PMMH chains are disposed to stammering if the likelihood estimator (12) has high variability. A chain moving into a particular state on the basis of an unusually large likelihood estimate will tend to remain there for a prolonged period, resisting even very local proposals; it will finally concede to a new state after a perhaps inordinate amount of time. The source of variability is both sampling and resampling error in the particle filter [27, 16]. The methods proposed in this work target a reduction of the former for any given number of particles.

In Pitt et al. [28], one measure of performance used to compare algorithms is the standard deviation of repeated log-likelihood estimates at the true parameter value. The truth is not, of course, known for real observational data, but nevertheless it

is helpful to perform a few pilot runs at carefully selected $\boldsymbol{\theta}$ and tune M until the statistic is tolerable. While a satisfying configuration might be found for particular regions, rarely is one achievable across the full expanse of the prior: the estimator can be strongly heteroskedastic, and not independent of the parameters themselves! The problem is particularly acute when the process model itself is used as part of the proposal within the APF. Small values of those parameters which prescribe process noise variance, for example, may lead to narrow proposal distributions that amplify weight variance, and so the variability of likelihood estimates. The mixing properties of the model may also be affected by, for example, gradient- and decay-related parameters.

To explore this behaviour, we approximate, for any sample $\{\mathbf{x}_0, \boldsymbol{\theta}\}$, the long-term acceptance rate of repeated proposals at that point. We refer to this as the *conditional acceptance rate* (CAR) at $\{\mathbf{x}_0, \boldsymbol{\theta}\}$, with details given in §A. By computing this statistic at many points across the space of parameters and initial conditions, it is possible to explore the heteroskedasticity of the likelihood estimator. While related to the alternative metric of standard deviation [28], the CAR is more directly interpretable as to the impact of variability on the acceptance rate of a chain, and also accommodates any asymmetry in that variability. We prefer it for these reasons.

4 Proposal mechanisms for the collapsed state-space model

The combination of the auxiliary particle filter and collapsed state-space model permits the design of arbitrary proposal distributions, even for those models which would otherwise have no closed-form transition density. Three proposal schemes are considered in this work, each with and without a look-behind, to give six methods in total. The first scheme is the ordinary bootstrap particle filter [11]. The remaining two use the UKF in different ways, one similar to the existing unscented particle filter [37], albeit in noise space. All six methods are detailed in this section; consult Table 1 for the quintessentials.

4.1 Bootstrap particle filter

The simplest approach, that of the bootstrap filter [11], sets

$$q_t(\mathbf{u}_{1:t}) = p(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})p(\mathbf{u}_t)\,. \tag{15}$$

This is straightforward to apply, and indeed does not require a collapse to noise terms for tractability. Note that both factors of (15) cancel in the weight computation of (10), so that particles are merely weighted by likelihoods. We denote this method PF0.

Lacking $p(\mathbf{y}_t|\mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta})$, an analytical look-ahead is not forthcoming. A deterministic single-point pilot look-ahead, simulating with $\mathbf{u}_t = \mathbf{0}$, can offer improvement in some circumstances [18]. We denote such a method PF1:

$$\begin{aligned} q_t(\mathbf{u}_{1:t}) &= p(\mathbf{u}_{1:t-1}|\mathbf{u}_t = \mathbf{0}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})p(\mathbf{u}_t) && (16)\\ &\propto p(\mathbf{y}_t|\mathbf{u}_t = \mathbf{0}, \mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta})p(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})p(\mathbf{u}_t)\,. && (17)\end{aligned}$$

Method	Description	q_1	q_2	Number of propagations
PF0	Bootstrap particle filter, no look-behind	$p(\mathbf{u}_{1:t-1} \mid \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})$	$p(\mathbf{u}_t)$	M
PF1	Bootstrap particle filter, single-point pilot look-behind	$p(\mathbf{u}_{1:t-1} \mid \mathbf{u}_t = \mathbf{0}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$	$p(\mathbf{u}_t)$	$2M$
MUPF0	Marginal unscented particle filter, no look-behind	$p(\mathbf{u}_{1:t-1} \mid \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})$	$\hat{p}_{\mathcal{N}}(\mathbf{u}_t \mid \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$	$M + 2(N_u + N_x + N_y) + 1$
MUPF1	Marginal unscented particle filter, single-point pilot look-behind	$p(\mathbf{u}_{1:t-1} \mid \mathbf{u}_t = \hat{\boldsymbol{\mu}}_{\mathcal{N}}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$	$\hat{p}_{\mathcal{N}}(\mathbf{u}_t \mid \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$	$2M + 2(N_u + N_x + N_y) + 1$
CUPF0	Conditional unscented particle filter, no look-behind	$p(\mathbf{u}_{1:t-1} \mid \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})$	$\hat{p}_{\mathcal{N}}(\mathbf{u}_t \mid \mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_t)$	$2M(N_u + N_y + 1)$
CUPF1	Conditional unscented particle filter, with look-behind	$\hat{p}_{\mathcal{N}}(\mathbf{u}_{1:t-1} \mid \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$	$\hat{p}_{\mathcal{N}}(\mathbf{u}_t \mid \mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_t)$	$2M(N_u + N_y + 1)$

Table 1: Quick reference chart of the auxiliary particle filter proposal mechanisms explored. See §4 for details.

4.2 Marginal unscented particle filter

We next attempt to draw on analytical approximations to inform the proposal distribution. The particular focus is on the UKF, which, for modest state sizes, tends to outperform other approximate nonlinear Kalman filtering approaches [39], such as the extended and ensemble [6, 7] variants. The UKF approximates the time-marginals $p(\mathbf{u}_t, \mathbf{x}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$ using a Gaussian distribution. We denote the approximation $\hat{p}_{\mathcal{N}}(\mathbf{u}_t, \mathbf{x}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$ henceforth. At each time, $2(N_u + N_x + N_y) + 1$ number of σ-points are crafted about the mean of the Gaussian distribution, propagated through the process model and specifically weighted to compute $\hat{p}_{\mathcal{N}}(\mathbf{u}_t, \mathbf{x}_t, \mathbf{y}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})$. Conditioning this on the actual observed value $\mathbf{y}_t$ delivers the approximate time marginal $\hat{p}_{\mathcal{N}}(\mathbf{u}_t, \mathbf{x}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$. See Julier and Uhlmann [15] and Wan and van der Merwe [39] for details.

The first UKF-based approach adopted is to use the *marginal* UKF approximations $\hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$ at each time as a common proposal for each particle. We call this the *marginal unscented particle filter* (MUPF), and without look-ahead, specifically MUPF0:

$$q_t(\mathbf{u}_{1:t}) = p(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})\hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})\,. \tag{18}$$

We can again consider a deterministic single-point pilot look-ahead, although this time might assume that fixing $\mathbf{u}_t$ to $\hat{\boldsymbol{\mu}}_t$, the mean estimate obtained from the UKF, will provide a better result than fixing to $\mathbf{0}$. We denote this MUPF1:

$$\begin{aligned} q_t(\mathbf{u}_{1:t}) &= p(\mathbf{u}_{1:t-1}|\mathbf{u}_t = \hat{\boldsymbol{\mu}}_t, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})\hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t}) & (19)\\ &\propto p(\mathbf{y}_t|\mathbf{u}_t = \hat{\boldsymbol{\mu}}_t, \mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta})p(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})\hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})\,. & (20)\end{aligned}$$

Note that (18) and (20) use the same time-marginal $\hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$ for each particle, not the conditional for the mth particle, $\hat{p}^m_{\mathcal{N}}(\mathbf{u}^m_t|\mathbf{u}^m_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})$, which may be preferred conceptually (and is the basis for the next scheme, see §4.3). This is a computational concession, permitting the UKF to be run asynchronously to the particle filter (see §B.1), but is not without justification for fast-mixing models regardless. The overhead of the method is slight, with only $2(N_u + N_x + N_y) + 1$ additional propagations for each particle filter. This is likely small compared to M, the number of propagations of the particle filter, which typically scales exponentially with dimension.

In the case where $\mathbf{U}_{1:T} \sim$ i.i.d. $\mathcal{N}(0, 1)$, the also-Gaussian marginals of the UKF afford some efficiencies in the weight calculation after slotting (18) or (20) into (10). These are given in §B.2. This also suggests that, even in the presence of a closed-form transition density over state variables, a collapse of the state-space model to noise terms may still be of benefit, computationally, to speed up weight calculations.

4.3 Conditional unscented particle filter

Separate UKF look-aheads for each particle may be preferred to a marginal lookahead, particularly for slower-mixing models. We refer to this as the *conditional unscented particle filter* (CUPF). It is not dissimilar to the unscented particle filter [37], besides the collapse to noise terms. Without look-behind, the CUPF0 method uses:

$$q_t(\mathbf{u}_{1:t}) = p(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})\hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})\,. \tag{21}$$

Note that the proposals are now different for each particle, as, after drawing from the first component, a UKF must be run to construct the second component for sampling. The proposal is a substantially more expensive operation as a result. Each UKF requires $2(N_u + N_y) + 1$ number of σ-points. While this is fewer than the MUPF (as the starting state for each is the single point $\mathbf{x}_{t-1}^m$), the total number of propagations required is nonetheless substantially more, at $2M(N_u + N_y + 1)$. Furthermore, the UKFs must be synchronous to the PF, eliminating the opportunity for asynchronous execution on multiple devices as described for the MUPF methods in §B.1. As consolation, weights for the look-behind are essentially free, giving a CUPF1 method that potentially improves substantially on CUPF0 at little additional cost:

$$
\begin{aligned}
q_t(\mathbf{u}_{1:t}) &= \hat{p}_{\mathcal{N}}(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})\hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t}) && (22)\\
&\propto \hat{p}_{\mathcal{N}}(\mathbf{y}_t|\mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta})p(\mathbf{u}_{1:t-1}|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t-1})\hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{u}_{1:t-1}\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t})\,. && (23)
\end{aligned}
$$

For higher-dimensional models the complexity may be dominated by the computation of the Schur complement required in the Kalman update equations. This is simplified by the independence of noise terms, however, with the particular case for Gaussian noise treated in §B.2.

5 Case studies in marine biogeochemistry

The proposed methods are assessed empirically on two models in the domain of marine biogeochemistry. Comparisons include conditional acceptance rates (CAR) and the spatial dependence of these, overall acceptance rates and Markov chain convergence rates.

Methods are assessed in two configurations (see Table 2): *particle-matched*, where the number of particles, M, is the same for all methods, and *compute-matched*, where M is set for all methods so as to roughly equate the number of propagations performed, where these include look-ahead pilots and UKF σ-points. The compute-matched case is meant to control, to some extent, for the particular software implementation of each algorithm. Rather than matching wallclock execution time, it instead recognises that propagations typically, and almost universally, dominate the profile of any particular implementation (80-90% of execution time in the authors' experience).

Simulated data is used, generated from the case study models themselves. This has a number of advantages over real observational data: execution time can be managed to facilitate the many runs required for some diagnostics, the model is perfect in the generative sense, capturing all (and only) those processes influencing observations, and a known ground truth for all latent and observed variables is available for validation of the methods. Note immediately, however, that given the simulated data set has noisy and finite number of observations, a hypothetically perfect Bayesian inference will not recover the ground truth exactly: not only is the ground truth not the maximum likelihood estimate (MLE), but the prior may bias the result away from this.

An observational data set is available, specifically from Ocean Station Papa (OSP), to which the second model has previously been fit using similar, but simpler, methods [25]. While the methods presented here can be used on this, the story for this data set is as yet incomplete, particularly around rigorous handling of its sparsity

Case study	Method	M	*Particle-matched* Acceptance rate	M	*Compute-matched* Acceptance rate
PZ	PF0	64	.182 (.003)	384	.245 (.002)
	PF1	64	.189 (.003)	192	.233 (.002)
	MUPF0	64	.208 (.003)	376	.251 (.002)
	MUPF1	64	.213 (.003)	184	.243 (.002)
	CUPF0	64	.209 (.003)	64	.209 (.003)
	CUPF1	64	.214 (.003)	64	.214 (.003)
NPZD	PF0	64	.142 (.003)	1536	.276 (.002)
	PF1	64	.139 (.006)	768	.254 (.003)
	MUPF0	64	.165 (.003)	1504	.278 (.002)
	MUPF1	64	.167 (.007)	736	.263 (.003)
	CUPF0	64	.169 (.003)	64	.169 (.003)
	CUPF1	64	.177 (.005)	64	.177 (.005)

Table 2: Configuration of methods for case studies, with mean (and standard deviation) of resulting acceptance rates across 256 chains.

(up to seven weeks between observations, compared to daily for the simulated sets here). Until that story is complete, we have chosen not to draw on this data for testing purposes, but pick up on the point in discussion (§6). Besides this issue of sparsity, however, we believe that the second case study is quite representative of real-world cases, especially in its nonlinear functional forms.

5.1 PZ model

The first model considered is a variant of the Lotka-Volterra differential system [19, 38], specifically over the predator-prey relationship of zooplankton and phytoplankton in a marine environment. Previously treated with a PMMH-style sampler [14], the intent here is to plumb deeper into the behaviour of the algorithm, the presence of just two parameters providing an ideal opportunity to visualise dependence of CAR on $\mathbf{\Theta}$. This *PZ* (phytoplankton and zooplankton) model modifies the classic Lotka-Volterra with the addition of a quadratic mortality term for zooplankton and stochastic growth term for phytoplankton. The stochasticity admits varying growth rates in phytoplankton without explicitly modelling contributory factors such as light and temperature, and thus exemplifies how such uncertainties can be treated by the introduction of stochasticity into an otherwise deterministic model [14, 25].

The state of the model is given by $\mathbf{X}_t = \{P_t, Z_t, \alpha_t\}$, with P and Z denoting concentrations of phytoplankton and zooplankton, respectively, and α the stochastic growth rate of phytoplankton. These interact via:

$$\frac{dP}{dt} = \alpha_t P - cPZ \tag{24}$$

$$\frac{dZ}{dt} = ecPZ - m_l Z - m_q Z^2 \,. \tag{25}$$

Here, t is time in days, with prescribed constants $c = .25$, $e = .3$, $m_l = .1$ and $m_q = .1$. The stochastic growth term, α_t, is drawn daily as $\alpha_t \sim \mathcal{N}(\mu, \sigma)$. Parameters to be estimated are $\Theta = \{\mu, \sigma\}$. Uniform prior distributions are assigned to the parameters, $\mu \sim \mathcal{U}(0, 1)$ and $\sigma \sim \mathcal{U}(0, .5)$. Log-normal distributions are placed over the initial conditions, $\ln P \sim \mathcal{N}(\ln 2, .2)$ and $\ln Z \sim \mathcal{N}(\ln 2, .1)$. Phytoplankton ($P$) is observed with log-normal noise:

$$\ln P_{\text{obs}} \sim \mathcal{N}(\ln P, .2)\,. \tag{26}$$

Simulated data is used by integrating forward a single trajectory for 100 days, taking P daily and adding observation noise. Z is unobserved.

A joint UKF is first applied, state augmented to include parameters, obtaining a preliminary approximation to the posterior over these. This provides the starting distribution for Markov chains, as well the covariance (after scaling by 0.18) for a random-walk Gaussian proposal distribution. This scaling factor is chosen using pilot runs with the PF0 method at 64 particles. Starting at the rule-of-thumb $2.4^2/N_\theta = 2.88$ [9], it is halved (four times) until a mixing rate close to the rule-of-thumb 23% [9] is achieved in the first 500 steps of the chain. Initial conditions are importance sampled in the particle filter, i.e. the target is factorised according to (5). Mixing might be improved by using (6); the ulterior motive, already mentioned, is for ready visualisation of Markov chain behaviour over only two dimensions.

Conditional acceptance rates are computed (as described in §A) at 1024 points on a regular grid across the uniform prior distribution, using 200 likelihood evaluations at each point. To produce contours of the surface, a Gaussian process is fit to the points by maximum likelihood, with a constant mean function, isotropic squared exponential covariance function and Gaussian likelihood [31, 30]. Results are presented for the particle-matched case in Figure 2, and compute-matched case in Figure 3.

Multiple chains are run for each method, with the multivariate $\hat{R}^p$ statistic of Brooks and Gelman [2] computed at regular intervals to assess rapidity of convergence (Figure 4). The posterior distribution obtained over parameters is marked in Figures 2 & 3, using samples drawn across all chains for each method. For a select chain, the state posterior is visualised in Figure 5, along with the ground truth trajectory and observations for comparison.

5.2 NPZD model

The introduction of nutrients, N, and detritus, D, into the PZ model provides a more realistic system with which real observational data can begin to be assimilated: an NPZD model. These additional terms are accompanied by various environmental forcings and rate processes that produce a more challenging model, with nonlinear responses ranging from convergence, to periodicity, to chaos. The full details and motivation behind the model are given in Parslow et al. [25]. A brief description to elucidate some of the complexity is given in §C.

A data set is generated by simulating the model with artificial forcing for 100 days, from which nutrient (N) and chlorophyll-a ($Chla$) observations are produced daily.

A joint UKF followed by an unscented Rauch-Tung-Striebel smoother [URTSS, 32] are applied, taking the joint covariance over initial conditions and parameters (scaled

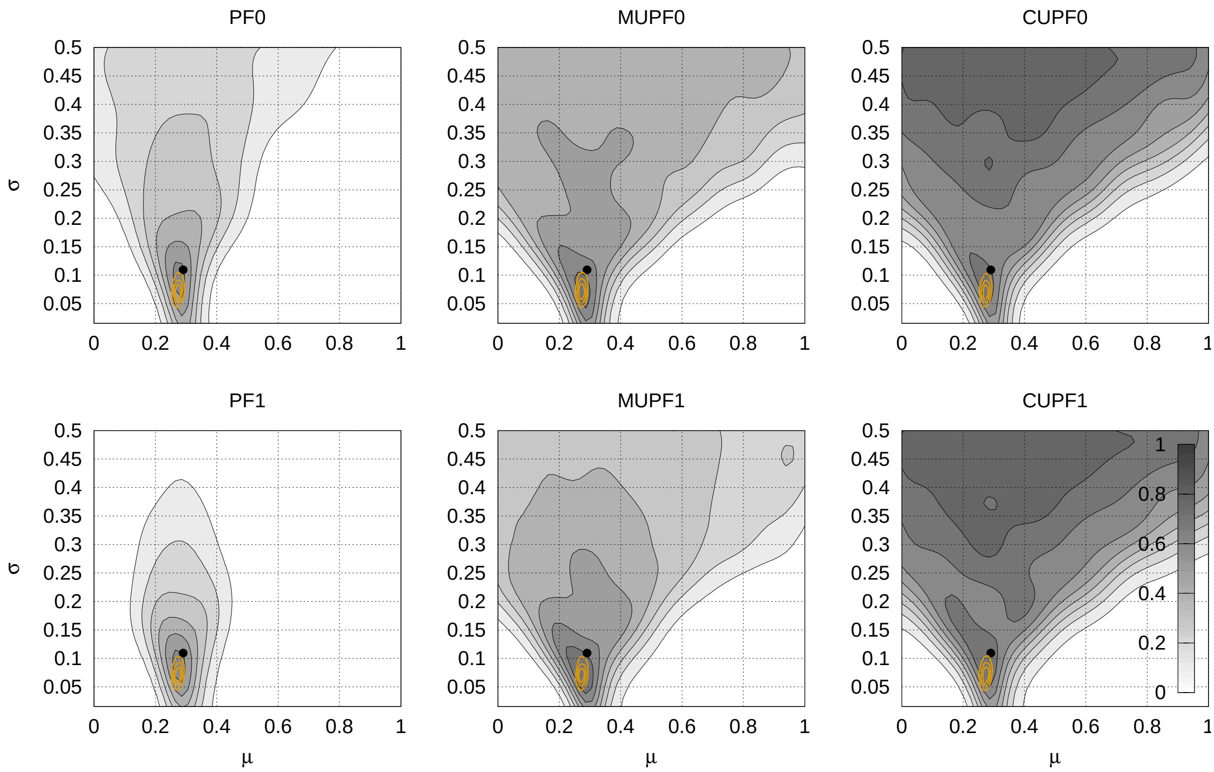

Figure 2: Conditional acceptance rates for each method applied to to the PZ case study, particle-matched, across the support of the uniform prior distribution. Darker shading denotes higher CAR (see key bottom right). The dots at approximately (0.3, 0.1) in each plot mark the ground truth parameters from which the data is simulated, and the bold contours nearby the posterior distribution obtained.

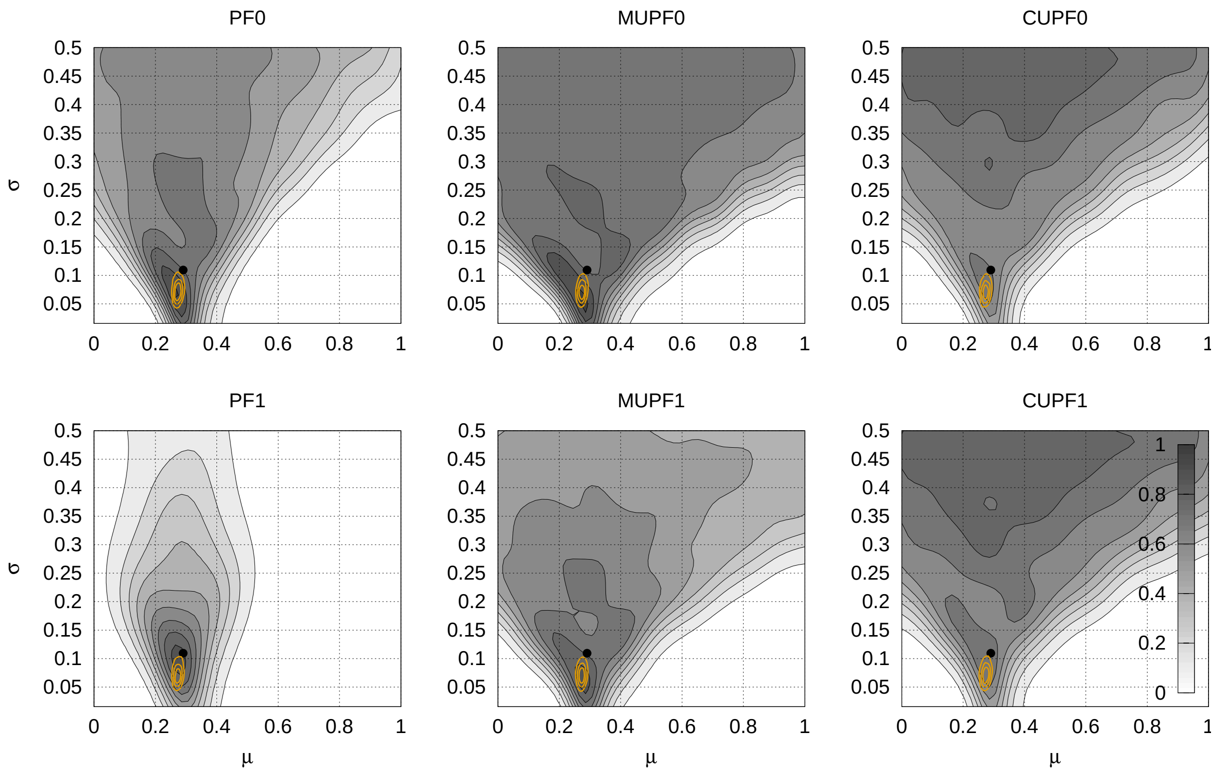

Figure 3: Conditional acceptance rates for each method applied to to the PZ case study, compute-matched. See Figure 2 caption for details.

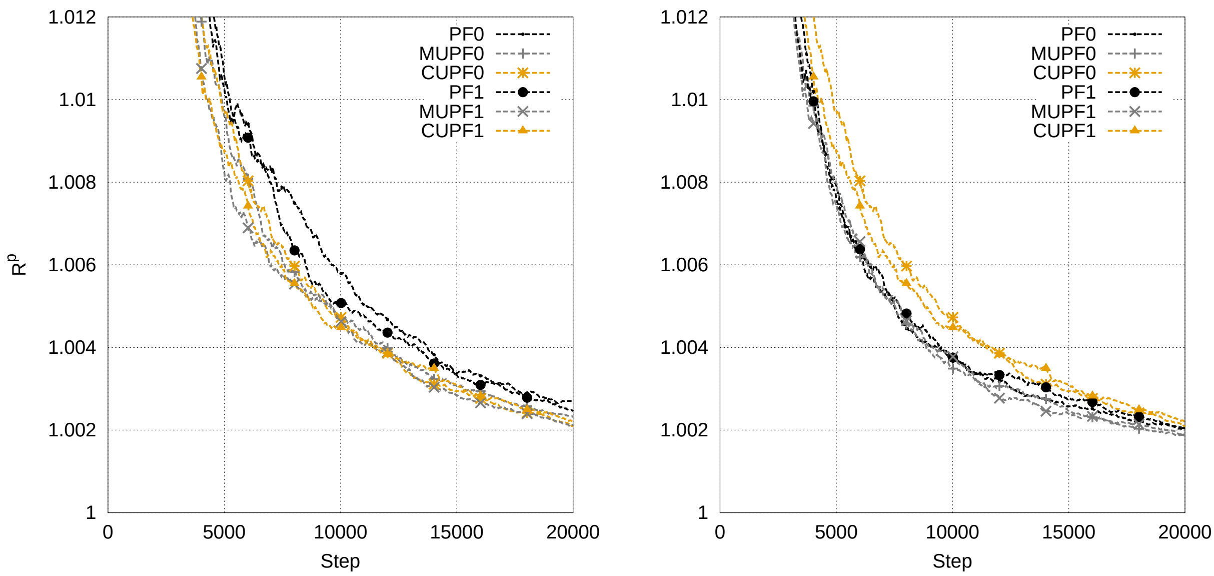

Figure 4: Convergence rates of Markov chains for the PZ case study, **(left)** particle-matched, and **(right)** compute-matched. Each line shows the evolution of the $\hat{R}^p$ statistic of Brooks and Gelman [2] for a particular method (see key), as the number of steps taken increases. The statistic is computed using 256 chains for each method. A more rapid approach to 1 indicates faster convergence.

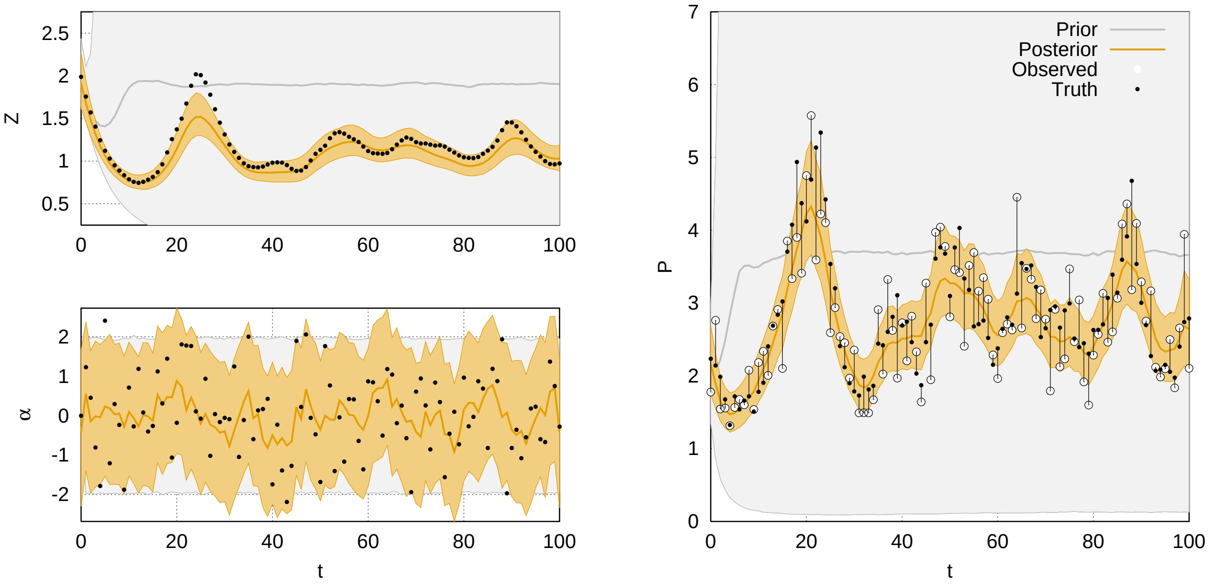

Figure 5: Time-marginal posteriors over state variables (P and Z) and noise term (α) for the PZ case study. Results are as obtained by PMMH using the CUPF1 method. Other methods give comparable results, although not shown. Bold lines of the prior and posterior denote the median, and shaded regions the 95% credibility interval, of the distribution.

by .012) for a random-walk Gaussian proposal for Markov chains, and both the joint mean and covariance as the starting distribution. Conditional acceptance rates are computed (as described in §A) at points drawn randomly from the prior distribution. Pairwise copulas between parameters and the CAR, computed empirically, are shown in Figure 6, with the empirical cdfs of those CAR in Figure 7. Figure 8 shows the covariance matrix over parameters obtained by both the PMMH and approximating URTSS method. The $\hat{R}^p$ statistic [2] is computed across multiple chains and shown in Figure 9. The univariate posterior marginal distributions of parameters and state are given in Figures 10 & 11.

6 Discussion

The collapse of a state-space model to its independent noise terms yields a closed-form transition density regardless of the functional forms of the process under study. Once this is acquired, the proposal distributions within an auxiliary particle filter (APF) can be tuned to improve sampling efficacy. It should be expected that the posterior densities over such noise terms may exhibit significant skew and kurtosis, for which it is difficult to devise good proposal distributions. Nevertheless, for the realistic marine biogeochemical models explored in this work, there is demonstrable utility in tuning merely the mean and covariance of a multivariate Gaussian proposal, increasing conditional acceptance rates, reducing variance in the likelihood estimator and improving overall convergence rates.

A common ill of the particle marginal Metropolis-Hastings (PMMH) sampler is

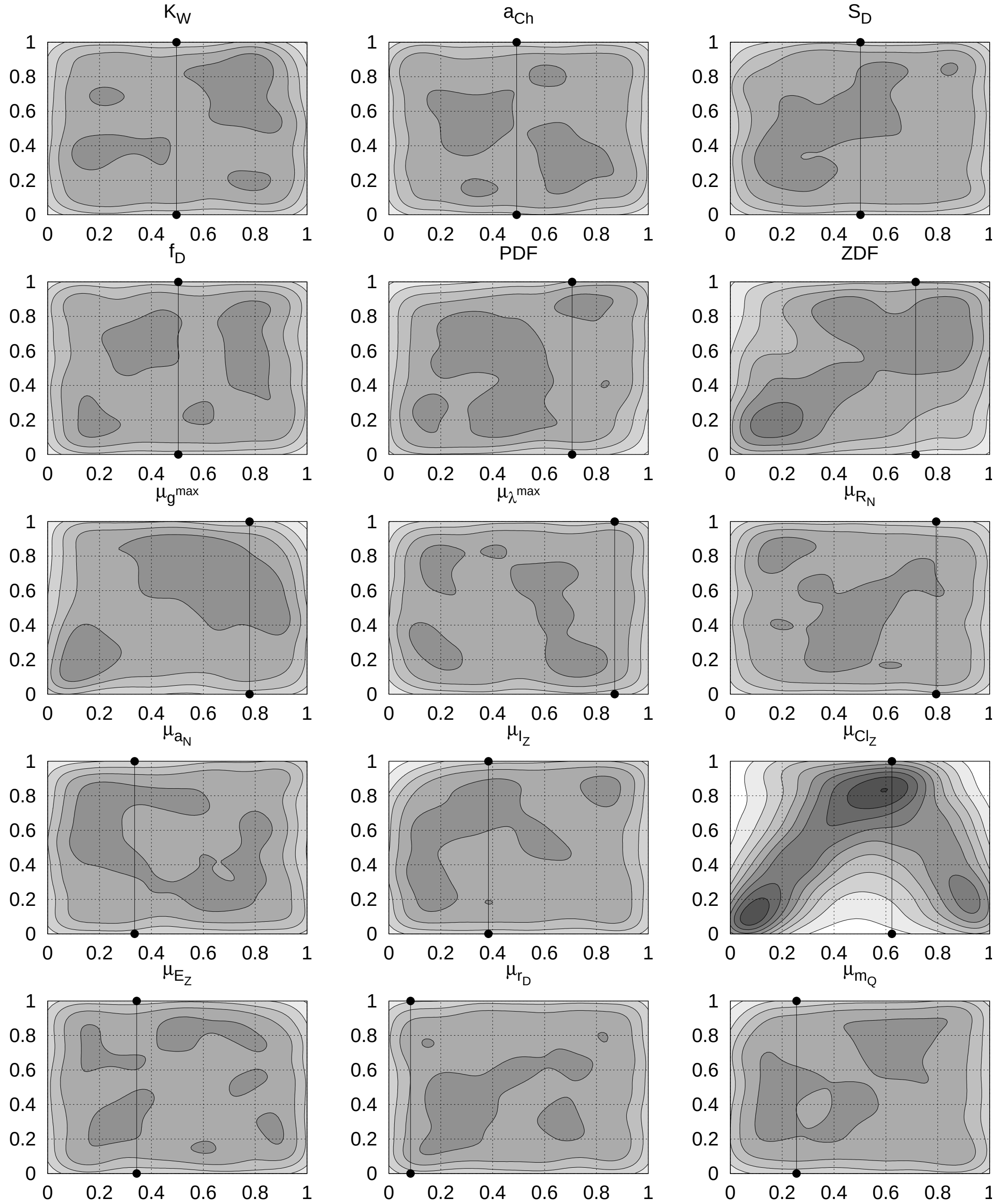

Figure 6: Estimated copula functions between parameters (x-axes) and CAR (y-axes) for the NPZD case study, using the CUPF1 method (other methods give similar results). For parameters, the prior univariate cumulative density functions (cdf) are used for transformation to uniform marginals; for the CAR, the empirical cdf is used. The copula function is approximated using a kernel density estimate of bandwidth .075 over 4096 points sampled from the prior distribution over parameters, with CARs computed from 200 likelihood evaluations at each point. Edge effects are an unfortunate artifact of the kernel density estimate. Most striking is the significant sensitivity of the CAR to the μ_{Cl_Z} parameter, explained in the text.

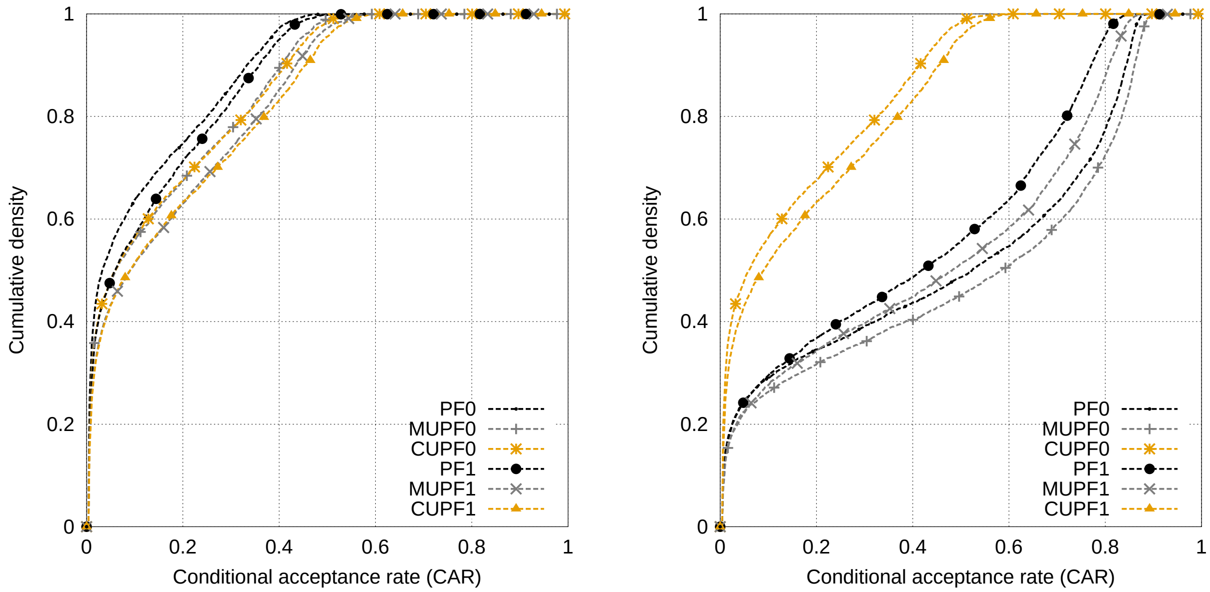

Figure 7: Empirical cdfs of the CAR for each method on the NPZD case study, **(left)** particle-matched, and **(right)** compute-matched. For each method, the CAR is computed at the same sample points used to construct Figure 6, then the empirical cdf over all of these CARs computed. As higher CARs are preferred, a lower cumulative density on the y-axis is preferred for any given rate on the x-axis.

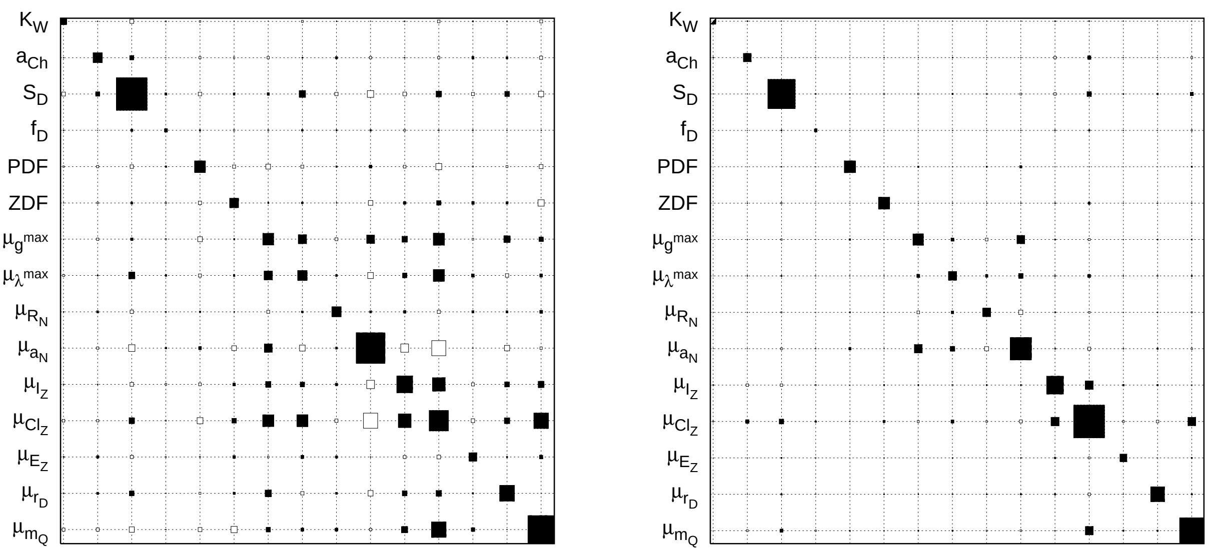

Figure 8: Covariance matrices over parameters of the NPZD model, as returned by **(left)** PMMH using CUPF1, and **(right)** URTSS. The area of each square is proportional to magnitude, with filled squares denoting positive, and empty squares negative, covariance. On the whole, some significant off-diagonal covariance is captured by the inexpensive URTSS method, from which a good proposal distribution might be constructed.

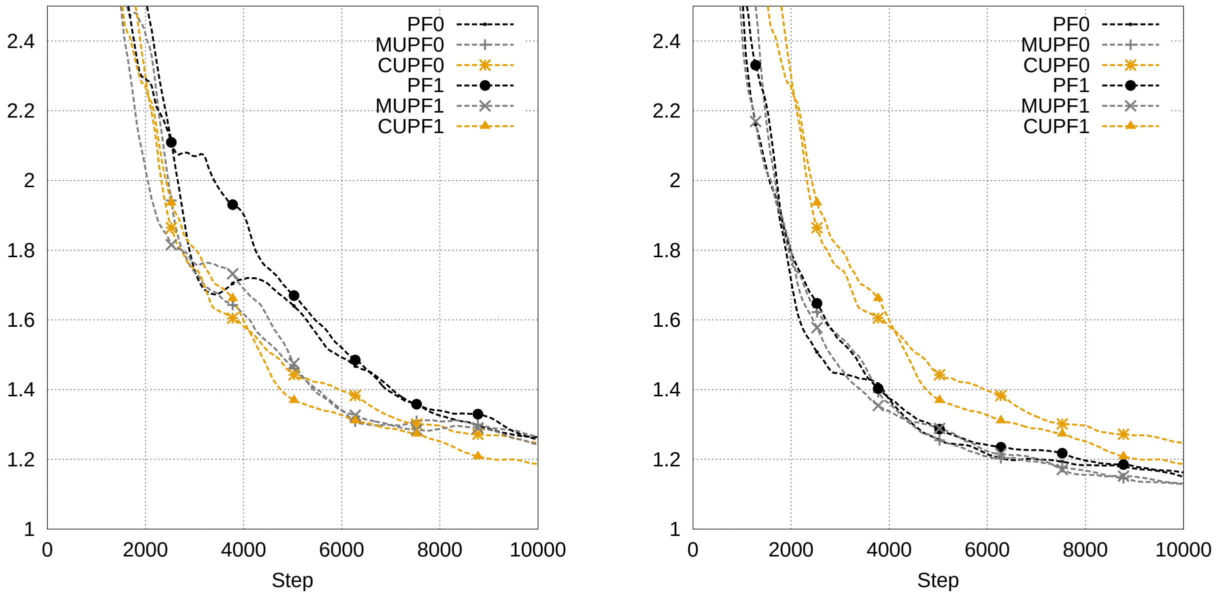

Figure 9: Convergence rates of Markov chains for the NPZD case study, **(left)** particle-matched, and **(right)** compute-matched. See Figure 4 for explanation. Computed using 256 chains for each method.

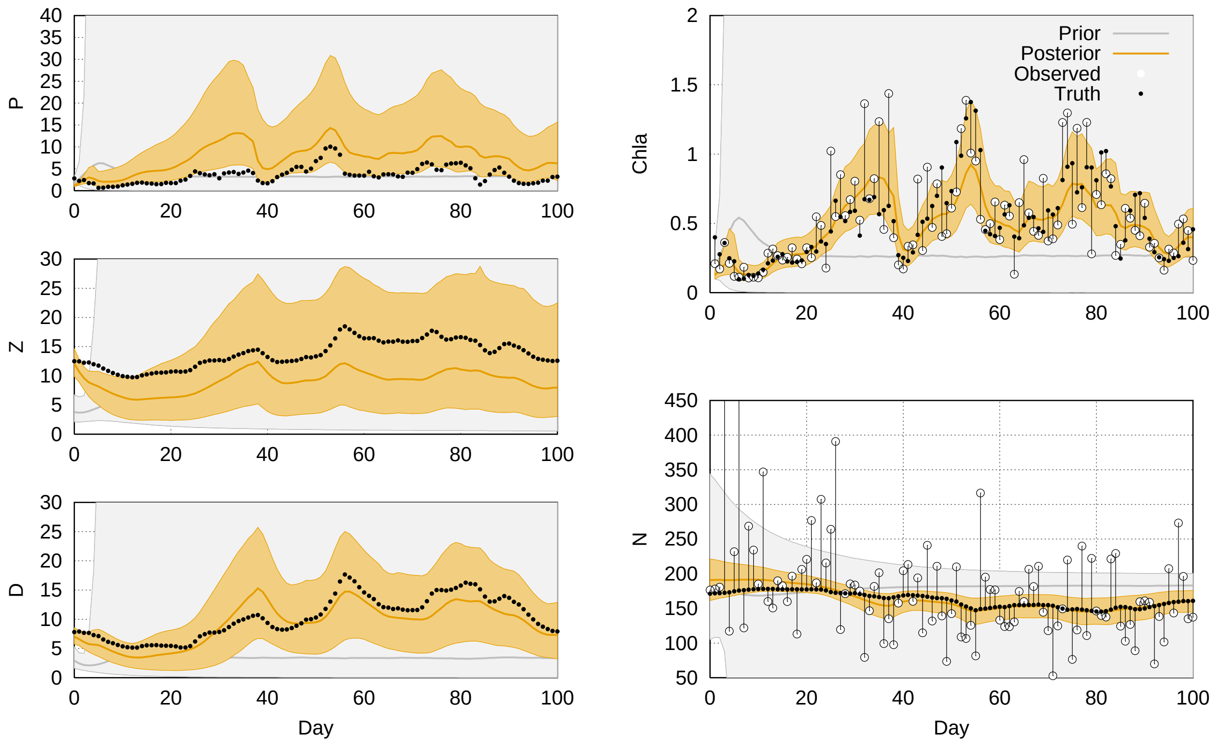

Figure 10: Time-marginal posterior distributions over state variables for the NPZD case study. Results are as obtained by PMMH using the CUPF1 method. Other methods give comparable results, although not shown. Bold lines of the prior and posterior denote the median, and shaded regions the 95% credibility interval, of the distribution.

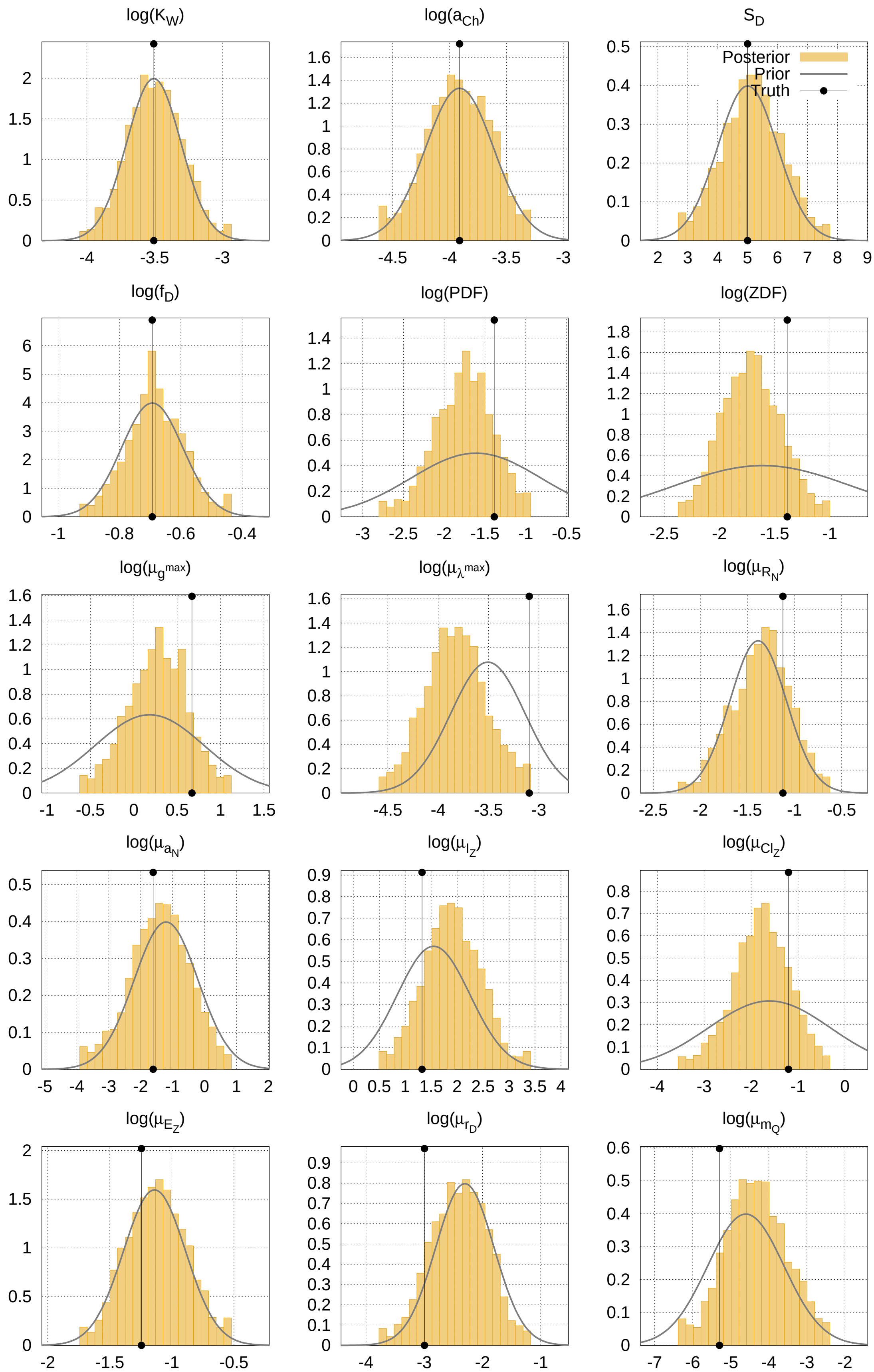

Figure 11: Prior and posterior distributions over parameters for the NPZD case study.

for it to enter a state on the basis of an unusually high likelihood estimate, and then remain there for a prolonged period. This may manifest a chain that stammers rather than mixing fluently. The behaviour is more pronounced when the variability of the likelihood estimator is high, and further confounded by heteroskedasticity across the domain of parameters. This work introduces the *conditional acceptance rate* (CAR) as a useful measure of the impact of such variability on the mixing of a PMMH chain at any given point. Applying this at multiple points reveals substantial variability in both the PZ (Figures 2 & 3) and NPZD (Figure 7) cases. For the PZ model, a clear decline in CAR away from the posterior mode is apparent (Figures 2 & 3), although larger values of the diffusive parameter σ lend improvement. The NPZD case study is more complex, many parameters having no apparent influence on CAR (Figure 6), at least over the support of the prior distribution. Again, however, there is some indication that larger values of diffusive parameters (specifically ZDF) improve CAR, which also falls away dramatically with distance from the posterior mode of a parameter to which the model is particularly sensitive, μ_{Cl_Z}[1].

The dependence of CAR on diffusive terms is not surprising given that the process model itself forms part of the APF proposal distribution in all methods explored. Broader proposals tend to deliver more effective importance samples, or are at least more forgiving of deviation between the modes of the proposal and target distributions. Note that the CAR will likely decline again if diffusive parameters become too large, but the prior distribution in both the PZ and NPZD cases constrains support to regions where a simple correlation is apparent. Because the prior distribution over parameters does not factor into any calculations made by the particle filter, the interpretation of the other apparent dependency of CAR should be distance from the maximum likelihood estimate (MLE), rather than distance from the posterior mode, or maximum *a posterior* (MAP) estimate.

Given the variability of the CAR, there are two potential pitfalls to avoid when using a PMMH sampler:

(a) initialisation in a region where CAR is low, and

(b) a particularly informative prior distribution that biases the posterior sufficiently far from the MLE to be in a region where CAR is low.

Either case may result in a failure of the Markov chain to converge, or even accept, in reasonable time. These should be considered failure modes of the PMMH sampler in much the same way as strongly correlated variables can cause slow mixing in the Gibbs sampler, or multiple modes can cause quasiergodicity in any MCMC procedure. The CAR is one avenue to diagnose this behaviour. Note that one cannot simply reduce the size of the proposal to arbitrarily increase acceptance rates, as with conventional MCMC with exact likelihoods. The construction of the CAR is as an estimate of the acceptance rate of repeated Dirac proposals at a single point of interest: the lower

[1]This parameter dictates the mean of the stationary distribution of the zooplankton clearance rate autoregressive. At low clearance rates, phytoplankton (P) will periodically escape zooplankton (Z) grazing control and begin a rapid bloom, triggering spikes in chlorophyll-a ($Chla$) that cannot be reconciled with observations. At high clearance rates, phytoplankton is relentlessly suppressed by zooplankton predation, keeping chlorophyll-a at much lower values than those observed.

limit of proposal size. Neither is relying on an arbitrary initialisation, and expecting a PMMH chain to converge to the target distribution, likely to succeed: even a smooth likelihood surface will appear noisy to this sampler. The aim should be to constrain variability to a tolerable level within the support of the posterior distribution, and use some approximating method to initialise the Markov chain within sight of this. The use of the UKF and URTSS for this purpose appears successful for the case studies here. As an aside, the approximate posterior covariance output from these methods seems to capture some significant off-diagonal elements, so as to make a more useful proposal covariance than a contrived orthogonal structure (Figure 8).

When the number of particles (M) is matched across methods, the MUPF and CUPF methods outperform the basic PF methods in both CAR (Figure 2, left-hand side of Figure 7) and convergence rate (left-hand side of Figures 4 & 9). We stress again, however, that the process model must mix sufficiently quickly for the proposal over noise terms to be of any benefit. This is particularly the case for MUPF, which neglects information on the current state of each particle. For slow-mixing models with long memory, the look-behind is far more important, and the tuning of proposals over noise terms may be additional computation for naught. Conversely, the look-behind is precisely only of benefit for such slow-mixing models with some memory. Relative to the PZ model, the NPZD model fits this description. For fast-mixing models the look-behind is of little benefit, and even harmful if the single-point pilot does not represent the whole conditional transition density well. This appears to be the case with the PZ model, where PF1 and MUPF1 exhibit smaller CAR than their PF0 and MUPF0 partners, exaggerated as the pilot becomes less representative at higher σ. Note that CUPF1 implements its lookahead using UKFs conditioned on each particle, not using single-point pilots, and this should rarely harm results. Indeed, CUPF1 requires comparably little additional computation over CUPF0, and is almost certainly worth using over it for additional robustness.

In compute-matched scenarios, the MUPF methods appear to retain some superiority (Figure 3, right-hand side of Figures 4, 7 & 9). The CUPF methods do not. We should hesitate to dismiss the conditional approach, however, noting that the computational burden is from repeated UKFs, which could be replaced with, say, extended Kalman filters (EKFs) to deliver similar returns for less computational investment. An interesting subplot arises from the CUPF methods in the PZ case study. No other method matches the CAR of CUPF0 and CUPF1 far from the posterior distribution (Figure 2), even after correction for compute time (Figure 3). This suggests that these methods may make a more robust choice for early steps in a Markov chain if a good initialisation is unavailable, later displaced by one of the cheaper methods once the posterior mode is in sight. Note, finally, that the overall acceptance rates achieved by each method are commensurate with their corresponding CARs in the vicinity of the posterior mode, rather than CARs across the prior as a whole (most clear in the PZ case, compare Table 2 against Figures 2 & 3). This is, of course, where a successful chain will spend the bulk of its time.

Finally, we return to practical matters on the marine biogeochemical case studies. At least one of the operational requirements of such models is their skill in prediction. Such a use typically requires slow-mixing process models, not the fast-mixing models desired for successful application of the MUPF and CUPF methods presented here.

We note, however, that proposals over noise terms provide some benefit for all but the slowest mixing of models. It is also worth noting that parameter estimation alone can significantly improve long-term predictions over those of the prior model, even for fast mixing models. The handling of sparse data is also an issue. With some extension, both the MUPF and CUPF methods could be made useful in this context, such as for the observational data sets available at the site of OSP. This might involve data augmentation [35], interleaving simulation of missing data values given the current state of the chain with a PMMH step conditioned on both real and imputed observations. This would mean that the distance to which the UKF (or EKF, or some other such method) looks ahead to the next observation (real or imputed) could be limited to an interval of time for which the approximation is credible. We have left this to future work.

The collapse of a state-space model to its noise terms seems a generally good approach to dealing with the absence of a closed-form transition density, which would otherwise prohibit the use of clever proposal distributions in inference. This work has demonstrated the utility of doing so to improve the acceptance and convergence rates of a PMMH sampler, working the UKF into the proposal mechanism of the inner particle filter. In doing so it has presented empirical results that elucidate some of the behaviours peculiar to the PMMH sampler, particularly highlighting the heteroskedasticity of the likelihood estimator, and the criticality of a good initialisation for the successful convergence of the Markov chain.

A Approximating the conditional acceptance rate (CAR)

At a particular point $\mathbf{z} = \{\mathbf{x}_0, \boldsymbol{\theta}\}$, we wish to compute the long-term acceptance rate of a particle marginal Metropolis-Hastings (PMMH) chain starting at $\mathbf{z}$, and continuously proposing an identity move to $\mathbf{z}$. A finite number, L, of likelihood estimates, $\hat{l}_1, \ldots, \hat{l}_L$, are made at $\mathbf{z}$ using some particle filtering method. From these a transition matrix $\mathbf{T} \in \mathbb{M}^{L \times L}$ is formed:

$$T_{ij} = \begin{cases} \frac{1}{L} \max\left[\frac{\hat{l}_j}{\hat{l}_i}, 1\right] & \text{if } i \neq j \\ 1 - \sum_{\substack{j=1 \\ j \neq i}}^{L} T_{ij} & \text{if } i = j\,. \end{cases} \tag{27}$$

In the extended space of the PMMH sampler, T_{ij} gives the probability of accepting a move to the jth sampled state, of likelihood $\hat{l}_j$, from the ith sampled state, of likelihood $\hat{l}_i$, under a uniform prior and proposal. From state i, the probability of accepting a uniformly drawn proposal is thus

$$\beta_i = 1 - T_{ii} + 1/L\,. \tag{28}$$

The $1/L$ term has the effect of counting a proposal to remain at state i as an acceptance.

From some arbitrary state, the Markov model defined by $\mathbf{T}$ may be run to equilibrium, where the probability of being in state i is simply the normalised term

$$\bar{l}_i = \frac{\hat{l}_i}{\sum_{j=1}^{L} \hat{l}_j}\,. \tag{29}$$

The long-term average acceptance rate, α, is then:

$$\alpha = \bar{\mathbf{l}} \cdot \boldsymbol{\beta}. \tag{30}$$

We call this the *conditional acceptance rate* at $\mathbf{z}$, abbreviated CAR.

The statistic may be computed more readily when the normalised likelihoods, $\bar{l}_1, \ldots, \bar{l}_L$, are pre-sorted in ascending order. In this case:

$$\beta_i = \frac{1}{L}\left[\sum_{j=1}^{i} \frac{\bar{l}_j}{\bar{l}_i} + L - i\right]. \tag{31}$$

Proceeding from (30), let $\mathbf{c}$ be the vector of inclusive prefix-sums over $\bar{\mathbf{l}}$ (i.e. $c_i = \sum_{j=1}^{i} \bar{l}_j$), and proceed:

$$\begin{aligned}
L\alpha &= \sum_{i=1}^{L} \bar{l}_i \left[\sum_{j=1}^{i} \frac{\bar{l}_j}{\bar{l}_i} + L - i\right] && (32)\\
&= \sum_{i=1}^{L}\sum_{j=1}^{i} \bar{l}_j + \sum_{i=1}^{L} \bar{l}_i (L - i) && (33)\\
&= \sum_{i=1}^{L} c_i + \sum_{i=1}^{L} c_i - \sum_{i=1}^{L} \bar{l}_i && (34)\\
&= 2\sum_{i=1}^{L} c_i - 1\,. && (35)
\end{aligned}$$

Thus, each α is readily computed by sorting and normalising $\hat{\mathbf{l}}$ to produce $\bar{\mathbf{l}}$, prefix-summing over this to compute $\mathbf{c}$, and summing over this, combined with some simple scalar operations, to close.

B Computational matters

B.1 Asynchronity in the marginal unscented particle filter

For the MUPF methods, observe that while the PF estimate of $p(\mathbf{u}_t|\mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta})$ depends on prior evaluation of the UKF estimate $\hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{u}_{1:t-1}, \mathbf{x}_0, \boldsymbol{\theta})$ to form a proposal distribution, there is no reciprocal dependence of the UKF on the PF. Consequently, the UKF may be run asynchronously to the PF, as long as it remains at least one step ahead at all times. This is not the case for CUPF, which requires the samples $\mathbf{u}_{1:t-1}^{1:M}$ from the PF before forming the proposal distributions for time t.

This property is particularly advantageous in a heterogeneous hardware context. The computational expense of repeated particle filters in PMCMC is offset by the inherent parallelism of the SMC component. The large number of independent particle propagations is particularly amenable to implementation on graphics processing units [GPUs, 17, 23, 22], being characterised by SIMD (Single-Instruction Multiple-Data) vector operations. The UKF, on the other hand, is dominated by matrix operations such as the Cholesky factorisation, for which GPU implementations [20] tend to

be favourable over central processing unit (CPU) implementations only for very many dimensions [36]. For a moderate number of dimensions, on the order of a few hundred, CPU execution of the UKF is preferred. If UKF execution time on the CPU is less than PF execution time on the GPU, MUPF may be employed at negligible overhead to the ordinary bootstrap particle filter, assuming that the CPU would otherwise idle during GPU execution.

B.2 Weight computations for Gaussian noise and proposal

Consider the weight computation for the auxiliary particle filter in (10), and in particular the factor $p(\mathbf{u}_t)/q_t(\mathbf{u}_t)$, where $p(\mathbf{u}_t) \equiv \text{i.i.d.}\,\mathcal{N}(0,1)$ and $q_t(\mathbf{u}_t) \equiv \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For models with Gaussian noise, this case includes the MUPF and CUPF methods introduced in §4.

To sample $\mathbf{u}_t^m$ from $q_2(\cdot)$ in (11), one draws $\boldsymbol{\xi}_t^m \sim \text{i.i.d.}\,\mathcal{N}(0,1)$ and sets $\mathbf{u}_t^m = \mathbf{L}\boldsymbol{\xi}_t^m + \boldsymbol{\mu}$, where $\mathbf{L}$ denotes the lower-triangular Cholesky factor of $\boldsymbol{\Sigma}$. Let

$$
\begin{aligned}
\ln \upsilon_t^m &= \ln p(\mathbf{u}_t) - \ln q_t(\mathbf{u}_t) && (36)\\
&= \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{u}_t^m)^T\mathbf{u}_t^m + \frac{1}{2}(\mathbf{u}_t^m - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{u}_t^m - \boldsymbol{\mu}) && (37)\\
&= \ln|\mathbf{L}| + \frac{1}{2}\left[(\boldsymbol{\xi}_t^m)^T\boldsymbol{\xi}_t^m - (\mathbf{u}_t^m)^T\mathbf{u}_t^m\right]. && (38)
\end{aligned}
$$

Substituting υ_t^m for $p(\mathbf{u}_t)/q_t(\mathbf{u}_t)$ in (10) gives the appropriate weighting for particle m.

In the case of the MUPF, where $q_t(\mathbf{u}_t) \equiv \hat{p}_{\mathcal{N}}(\mathbf{u}_t|\mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_{1:t}) \equiv \mathcal{N}(\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}_t, \boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}}_t)$, continued propagation of the UKF requires computation of the Cholesky factor $\mathbf{L} = \hat{\mathbf{L}}_t$ anyway, so that the additional operations required for the weight adjustment are comparably negligible: two dot products and a single vector multiply-and-add for each particle, and a single sum over the diagonal of $\hat{\mathbf{L}}_t$ to compute $\ln|\hat{\mathbf{L}}_t|$. These may, of course, be combined into matrix operations to perform the computations for all particles together.

For the CUPF, denoting $k = a_t^m$ for brevity, $q_t(\mathbf{u}_t^m) \equiv \hat{p}_{\mathcal{N}}(\mathbf{u}_t^m|\mathbf{u}_{1:t-1}^k, \mathbf{x}_0, \boldsymbol{\theta}, \mathbf{y}_t) \equiv \mathcal{N}(\boldsymbol{\mu} = \hat{\mathbf{u}}_t^k, \boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}}_t^k)$, and each UKF need only deliver $\hat{\boldsymbol{\mu}}_t^k$ and $\hat{\mathbf{L}}_t^k$ after conditioning on $\mathbf{y}_t$. Producing $\hat{\mathbf{L}}_t^k$ is particularly cheap by exploiting the fact that the uncorrected prediction of $\hat{\boldsymbol{\Sigma}}_t^k$ is always $\mathbf{I}$, giving the measurement update:

$$
\begin{aligned}
\hat{\boldsymbol{\Sigma}}_t^k &\leftarrow \hat{\boldsymbol{\Sigma}}_t^k - \hat{\mathbf{C}}_{uy}^k(\hat{\boldsymbol{\Sigma}}_y^k)^{-1}(\hat{\mathbf{C}}_{uy}^k)^T && (39)\\
&= \mathbf{I} - \hat{\mathbf{C}}_{uy}^k(\hat{\mathbf{L}}_y^k)^{-1}(\hat{\mathbf{L}}_y^k)^{-T}(\hat{\mathbf{C}}_{uy}^k)^T, && (40)
\end{aligned}
$$

where $\hat{\boldsymbol{\Sigma}}_y^k = \text{cov}(\mathbf{Y}_t, \mathbf{Y}_t|\mathbf{u}_{1:t-1}^k, \mathbf{x}_0, \boldsymbol{\theta})$, and $\hat{\mathbf{C}}_{uy}^k = \text{cov}(\mathbf{U}_t, \mathbf{Y}_t|\mathbf{u}_{1:t-1}^k, \mathbf{x}_0, \boldsymbol{\theta})$. Because the Cholesky factor of $\mathbf{I}$ is of course $\mathbf{I}$, that of the updated $\hat{\boldsymbol{\Sigma}}_t^k$ may be computed using N_y rank-1 Cholesky downdates to $\mathbf{I}$ over the columns of $(\hat{\mathbf{L}}_y^k)^{-T}(\hat{\mathbf{C}}_{uy}^k)^T$. For small N_y, this will be substantially quicker than the full computation and Cholesky factorisation of the updated $\hat{\boldsymbol{\Sigma}}_t^k$. It may be possible to further exploit $\hat{\boldsymbol{\Sigma}}_t^k = \mathbf{I}$ in this operation, such as using sparse formulations [3], but we are yet to explore the possibilities thoroughly.

C NPZD model specification

The NPZD model represents the interaction of nutrients (N), phytoplankton (P), zooplankton (Z) and detritus (D), in the common currency of nitrogen, within the surface mixed layer of a body of water. The surface waters are modelled as a single box, subject to exogenous environmental forcings such as available light, temperature and changes in mixed layer depth. A full derivation of the model, including sources for the prior distribution over parameters, can be found in Parslow et al. [25]; the essentials for the present work are summarised here. We have adopted a slightly simpler version of the model than that presented in Parslow et al. [25], whereby exogenous forcings are considered time-invariant.

For reference as components of the model are introduced, the model state is $\mathbf{X} = \{P, Z, D, N, g^{\max}, \lambda^{\max}, R_N, a_N, I_Z, Cl_Z, E_Z, r_D, m_Q\}$, and the parameters $\mathbf{\Theta} = \{K_W, a_{Ch}, S_D, f_D, \mu_{g^{\max}}, \mu_{\lambda^{\max}}, \mu_{R_N}, \mu_{a_N}, \mu_{I_Z}, \mu_{Cl_Z}, \mu_{E_Z}, \mu_{r_D}, \mu_{m_Q}, PDF, ZDF\}$.

C.1 Noise model

The model features nine noise terms, ξ_i for $i = 1, \ldots, 9$, each coupled to a univariate autoregressive process B_i. Four of these are phytoplankton-related, given by

$$B_i(t + \Delta t) = B_i(t) \cdot (1 - \Delta t/\tau_P) + (\mu_i + PDF \cdot \sigma_i \xi_i) \cdot \Delta t/\tau_P \,, \tag{41}$$

where Δt is a discrete time step (one day), μ_i a parameter to be estimated, PDF a common *diversity factor* to be estimated, σ_i a prescribed scaling factor, and τ_P a common characteristic time scale, also prescribed. The remaining five autoregressive processes are zooplankton-related, modelled using the same form, with ZDF and τ_Z replacing PDF and τ_P, respectively.

Each process represents a property of the phytoplankton (zooplankton) community, the species composition of which will change with time. Rather than model individual species, the phytoplankton (zooplankton) community is modelled collectively, with diversity factors PDF and ZDF scaling stochastic drivers used to model the changing influence of community composition.

The four phytoplankton processes are $\{g^{\max}, \lambda^{\max}, R_N, a_N\}$, and the five zooplankton processes $\{I_Z, Cl_Z, E_Z, r_D, m_Q\}$. Each is accompanied by its matching noise term amongst $\mathbf{U} = \{\xi_{g^{\max}}, \xi_{\lambda^{\max}}, \xi_{R_N}, \xi_{a_N}, \xi_{I_Z}, \xi_{Cl_Z}, \xi_{E_Z}, \xi_{r_D}, \xi_{m_Q}\}$, and mean parameter amongst $\{\mu_{g^{\max}}, \mu_{\lambda^{\max}}, \mu_{R_N}, \mu_{a_N}, \mu_{I_Z}, \mu_{Cl_Z}, \mu_{E_Z}, \mu_{r_D}, \mu_{m_Q}\}$.

C.2 State model

The equations governing interactions between the remaining state variables $\{P, Z, D, N\}$ are:

$$\frac{dP}{dt} = g \cdot P - gr \cdot Z + \frac{\kappa}{MLD} \cdot (BCP - P) \tag{42}$$

$$\frac{dZ}{dt} = E_Z \cdot gr \cdot Z - m \cdot Z \tag{43}$$

$$\frac{dD}{dt} = (1 - E_Z) \cdot f_D \cdot gr \cdot Z + m \cdot Z - r \cdot D - S_D \cdot \frac{D}{MLD} + \frac{\kappa}{MLD} \cdot (BCD - D) \tag{44}$$

$$\frac{dN}{dt} = -g \cdot P + (1 - E_Z) \cdot (1 - f_D) \cdot gr \cdot Z + r \cdot D + \frac{\kappa}{MLD} \cdot (BCN - N)\,. \tag{45}$$

Here, g is the phytoplankton specific growth rate (per day, or d^{-1}), gr is the zooplankton specific grazing rate (mg P grazed per mg Z d^{-1}), m is the zooplankton specific mortality rate (d^{-1}), and r is the specific breakdown rate of detritus (d^{-1}). The fraction E_Z (a parameter to be estimated) of zooplankton ingestion is converted to zooplankton growth and, of the remainder, a fraction f_D allocated to detritus and the rest released as dissolved inorganic nutrient, N. The grazing rate gr, mortality rate m and growth rate g are not only functions of the state, but also prescribed exogenous forcings and physiological constants, described below.

C.3 Rate processes

A multiplicative temperature correction Tc is applied to all rate processes, for which a Q_{10} formulation for dependence on temperature, T, is used:

$$Tc = Q_{10}^{(T - T_{\mathrm{ref}})/10}\,, \tag{46}$$

where T_{ref} is a reference temperature, and Q_{10} a prescribed constant.

The zooplankton grazing rate (gr) is dependent on the phytoplankton concentration (zooplankton functional response):

$$gr = \frac{Tc \cdot I_Z \cdot A^{\upsilon}}{(1 + A^{\upsilon})}\,, \tag{47}$$

where υ is a given power; the relative availability of phytoplankton, A, is

$$A = \frac{Cl_Z \cdot P}{I_Z}\,; \tag{48}$$

I_Z is the maximum zooplankton ingestion rate (mg P per mg Z per day); Cl_Z is the maximum clearance rate (volume in m^3 swept clear per mg Z per day). For $\upsilon = 1$, (47) takes the form of a Type-2 functional response (standard rectangular hyperbola) [13], and for $\upsilon > 1$ a Type-3 sigmoid functional response.

A quadratic formulation for zooplankton mortality is adopted after Steele [33] and Steele and Henderson [34]:

$$m = Tc \cdot m_Q \cdot Z\,, \tag{49}$$

where the quadratic mortality rate m_Q has units of $\mathrm{d}^{-1}(\mathrm{mg}Z\mathrm{m}^{-3})^{-1}$. The detrital remineralization rate is dependent only on temperature:

$$r = Tc \cdot r_D \,, \tag{50}$$

where r_D prescribes the remineralisation rate at a reference temperature.

The phytoplankton specific growth rate, g, depends on temperature, T, available light or irradiance, E, and dissolved inorganic nutrient, N. It is expressed in terms of a maximum specific growth rate at the reference temperature, $g^{\max}$ (d^{-1}), a light-limitation factor, h_E, and a nutrient-limitation factor, h_N:

$$g = Tc \cdot g^{\max} \cdot h_E \cdot h_N/(h_E + h_N) \,. \tag{51}$$

The light-limitation factor is given by

$$h_E = 1 - \exp(-\alpha \cdot \lambda^{\max} \cdot E/g^{\max}) \,, \tag{52}$$

where α is the initial slope of the photosynthesis versus irradiance curve (mg C mg $Chla^{-1}$ mol photon^{-1} m^2), and $\lambda^{\max}$ is the maximum $Chla : C$ ratio (mg $Chla$ mg C^{-1}). Here, α is calculated as the product of the chlorophyll-specific absorption coefficient for phytoplankton, a_{Ch} (m^2 mg $Chla^{-1}$), and the maximum quantum yield for photosynthesis, Q (mg C mol photons^{-1}). E is the mean photosynthetic available radiation (PAR) in the mixed layer and is given by

$$E = E_0.(1 - \exp(-Kz))/Kz, \tag{53}$$

where E_0 is the mean daily photosynthetically available radiation (PAR) just below the air-sea interface and Kz is given by

$$Kz = (K_W + a_{Ch} \cdot Chla) \cdot MLD. \tag{54}$$

In (54), K_W is attenuation due to the seawater and a_{Ch}.

The nutrient-limitation factor is given by

$$h_N = \frac{N}{(g^{\max} \cdot Tc/a_N) + N} \,, \tag{55}$$

where a_N is the maximum specific affinity for nitrogen uptake (d^{-1} mg N^{-1} m^3).

The phytoplankton $N : C$ ratio, χ, predicted by the model is given by

$$\chi = \frac{\chi^{\min} \cdot h_E + \chi^{\max} \cdot h_N}{h_E + h_N} \,, \tag{56}$$

where $\chi^{\min}$ and $\chi^{\max}$ are the prescribed minimum and maximum $N : C$ ratios (mg N mg C^{-1}).

C.4 Boundary conditions

The simple one-box mixed layer model adopted here needs to allow for the effects of physical exchanges between the mixed layer and the underlying water mass. With the exception of BCN, all boundary conditions (BCP, BCD, BCZ) are set to zero for the experiments in this work. The variable κ sets the strength of the mixing; in this study, we assume that the lower two metres of the mixed layer are replenished daily with water from the sub-mixed layer. MLD and BCN are in this case time invariant and set to 40 m and 200 mg N m^3 respectively.

C.5 Observation model

The model predicts the phytoplankton $Chla : C$ ratio λ, and this can be combined with the $N : C$ ratio χ to convert phytoplankton biomass P (mg N m^{-3}) to a predicted $Chla$ concentration:

$$Chla = P \cdot (\lambda^{\max}/\chi^{\max}) \cdot h_N \cdot Tc/(R_N \cdot h_E + h_N)\,. \tag{57}$$

Both N and $Chla$ are observed, each with log-normal noise of 40%, i.e. $\ln N_{\text{obs}} \sim \mathcal{N}(\ln N, .4)$, and $\ln Chla_{\text{obs}} \sim \mathcal{N}(\ln Chla, .4)$. Observations can thus be written as $\mathbf{Y} = \{N_{\text{obs}}, Chla_{\text{obs}}\}$.

References

[1] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B*, 72:269–302, 2010.

[2] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

[3] T. A. Davis and W. W. Hager. Modifying a sparse Cholesky factorization. *SIAM Journal on Matrix Analysis and Applications*, 20:606–627, 1999.

[4] P. Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

[5] A. Doucet, N. de Freitas, K. Murphy, and S. Russel. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 176–183, 2000.

[6] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *Journal of Geophysical Research-Oceans*, 99 (C5):10143–10162, 1994. ISSN 0148-0227.

[7] G. Evensen. The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems Magazine*, 29:83–104, 2009.

[8] P. Fearnhead, O. Papaspiliopoulos, and G. O. Roberts. Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society Series B*, 70:755–777, 2008.

[9] A. Gelman, W. Gilks, and G. Roberts. Efficient Metropolis jumping rules. Technical Report 94-10, University of Cambridge, 1994.

[10] A. Golightly and D. Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52:1674–1693, 2008. doi: doi:10.1016/j.csda.2007.05.019.

[11] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140:107–113, 1993.

[12] W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[13] C. S. Holling. The functional response of predators to prey density and its role in mimicry and population regulation. *Memoirs of the Entomology Society of Canada*, 45:60, 1966.

[14] E. Jones, J. Parslow, and L. M. Murray. A Bayesian approach to state and parameter estimation in a phytoplankton-zooplankton model. *Australian Meteorological and Oceanographic Journal*, 59(SP):7–16, 2010.

[15] S. J. Julier and J. K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *The Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls, Multi Sensor Fusion, Tracking and Resource Management*, 1997.

[16] A. Lee. Towards smooth particle filters for likelihood estimation with multivariate latent variables. Master's thesis, University of British Columbia, 2008.

[17] A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 19:769–789, 2010. doi: doi:10.1198/jcgs.2010.10039.

[18] M. Lin, R. Chen, and J. S. Liu. Lookahead strategies for sequential Monte Carlo. Technical report, Rutgers University, Peking University and Harvard University, 2009.

[19] A. Lotka. *Elements of physical biology*. Williams & Wilkins, 1925.

[20] H. Ltaief, S. Tomov, R. Nath, and J. Dongarra. Hybrid multicore Cholesky factorization with multiple GPU accelerators. *IEEE Transaction on Parallel and Distributed Systems*, 2010.

[21] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[22] L. M. Murray. GPU acceleration of the particle filter: The Metropolis resampler. In *DMMD: Distributed machine learning and sparse representation with massive data sets*, 2011.

[23] L. M. Murray. GPU acceleration of Runge-Kutta integrators. *IEEE Transactions on Parallel and Distributed Systems*, 23:94–101, 2012.

[24] L. M. Murray and A. Storkey. Particle smoothing in continuous time: A fast approach via density estimation. *IEEE Transactions on Signal Processing*, 59:1017–1026, 2011. doi: 10.1109/TSP.2010.2096418.

[25] J. Parslow, N. Cressie, E. P. Campbell, E. Jones, and L. M. Murray. Bayesian learning and predictability in stochastic nonlinear models. Technical Report 865, Department of Statistics, The Ohio State University, 2012.

[26] M. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94:590–599, 1999.

[27] M. K. Pitt. Smooth particle filters for likelihood evaluation and maximisation. Technical Report 651, The University of Warwick, Department of Economics, July 2002.

[28] M. K. Pitt, R. S. Silva, P. Giordani, and R. Kohn. Auxiliary particle filtering within adaptive Metropolis-Hastings sampling. 2011. URL `http://arxiv.org/abs/1006.1914`.

[29] V. Plagnol and S. Tavar. Approximate Bayesian computation and MCMC. In *Monte Carlo and Quasi-Monte Carlo Methods*, 2003.

[30] C. E. Rasmussen and H. Nickisch. GPML Matlab code. Online, 2011. http://www.gaussianprocess.org/gpml/code/.

[31] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[32] S. Särkkä. Unscented Rauch-Tung-Striebel smoother. *IEEE Transactions on Automated Control*, 53:845–849, 2008.

[33] J. Steele. Role of predation in ecosystem models. *Marine Biology*, 35(1):9–11, 1976. ISSN 0025-3162.

[34] J. Steele and E. Henderson. The role of predation in plankton models. *Journal of Plankton Research*, 14(1):157–172, 1992.

[35] M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

[36] S. Tomov, R. Nath, H. Ltaief, and J. Dongarra. Dense linear algebra solvers for multicore with GPU accelerators. In *Proceedings of IPDPS'10*, 2010.

[37] R. van der Merwe, A. Doucet, N. de Freitas, and E. Wan. The unscented particle filter. *Advances in Neural Information Processing Systems*, 13, 2000.

[38] V. Volterra. *Animal Ecology*, chapter Variations and fluctuations of the number of individuals in animal species living together. McGraw-Hill, 1931. Translated from 1928 edition by R.N. Chapman.

[39] E. A. Wan and R. van der Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of IEEE Symposium on Adaptive Systems for Signal Processing Communications and Control*, pages 153–158, 2000.